



TECHNICAL REPORT TR-08-20-2009-CS-MU

Motion Refinement Based Progressive Side-Information Estimation for
Wyner-Ziv Video Coding ¹

Wei Liu, Lina Dong and Wenjun Zeng

Department of Computer Science
University of Missouri
Columbia, MO 65211
<http://www.cs.missouri.edu/~zeng/>

Aug. 20, 2009

Abstract

The accuracy of motion estimation (ME) plays an important role in improving the coding efficiency of Wyner-Ziv video coding (WZVC). Most existing WZVC schemes perform ME at the decoder side. The unavailability of the current frame at the decoder side usually limits the accuracy of ME, which results in the degradation of the side information (SI) quality, as well as the coding efficiency of WZVC. To improve ME accuracy, some works in the literature propose to decode the current frame progressively, and the decoder iteratively refines the motion field based on each partially-decoded image. In this paper, we present an analytical model to estimate the potential gain by employing multi-resolution motion refinement (MRMR), assuming the current frame is progressively decoded in the resolution dimension. Our theoretical results show that at high rates, MRMR outperforms motion extrapolation by as much as 5 dB in terms of SI quality, while falling behind conventional inter-frame ME by about 1.5 dB. In addition, by observing that decoder-side ME does not suffer from the bit-rate overhead in transmitting the motion, we investigate the performance gain when incorporating fractional-pel motion search, block matching with finer block sizes and multiple hypothesis prediction into the MRMR framework. Theoretical analysis shows significant gain provided by these techniques. A practical SI estimator is implemented which provides prediction performance comparable to that of H.264/AVC.

¹ This work is supported in part by NSF grant CNS-0423386. This work is also under review for publication in *IEEE Trans. Circuits & Systems for Video Technologies*.

Motion Refinement Based Progressive Side-Information Estimation for Wyner-Ziv Video Coding

Wei Liu, Lina Dong and Wenjun Zeng

Abstract—The accuracy of motion estimation (ME) plays an important role in improving the coding efficiency of Wyner-Ziv video coding (WZVC). Most existing WZVC schemes perform ME at the decoder side. The unavailability of the current frame at the decoder side usually limits the accuracy of ME, which results in the degradation of the side information (SI) quality, as well as the coding efficiency of WZVC. To improve ME accuracy, some works in the literature propose to decode the current frame progressively, and the decoder iteratively refines the motion field based on each partially-decoded image. In this paper, we present an analytical model to estimate the potential gain by employing multi-resolution motion refinement (MRMR), assuming the current frame is progressively decoded in the resolution dimension. Our theoretical results show that at high rates, MRMR outperforms motion extrapolation by as much as 5 dB in terms of SI quality, while falling behind conventional inter-frame ME by about 1.5 dB. In addition, by observing that decoder-side ME does not suffer from the bit-rate overhead in transmitting the motion, we investigate the performance gain when incorporating fractional-pel motion search, block matching with finer block sizes and multiple hypothesis prediction into the MRMR framework. Theoretical analysis shows significant gain provided by these techniques. A practical SI estimator is implemented which provides prediction performance comparable to that of H.264/AVC.

Index Terms—Wyner-Ziv video coding, motion estimation, multi-resolution processing, rate-distortion analysis.

I. INTRODUCTION

IN the last two decades, video coding has achieved a tremendous success using the hybrid coding structure, i.e. inter-frame motion compensated prediction (MCP) combined with intra-frame transform coding. Such a structure relies on a powerful encoder that performs heavy computation, while decoders can be relatively lightweight, working in a slave mode. In recent years, with the rapid development of wireless sensor networks, distributed video coding (DVC) [3] has become a hot research topic. The scenario in which DVC is

used often involves multiple senders (encoders) with limited processing capability and power supply, and a relatively powerful receiver (decoder). Therefore the computational burden needs to be shifted to the receiver. DVC provides a useful alternative and complement to the conventional hybrid coding, especially in the so-called “uplink” video transmission applications.

DVC is theoretically based on distributed source coding (DSC). The history of DSC dates back to Slepian and Wolf’s seminal work in 1973 [4]. It is proved that there is no performance loss to losslessly compress two statistically dependent sources even if the two encoders do not communicate with each other. Lossless DSC is usually referred to as Slepian-Wolf coding (SWC). The relationship between SWC and channel coding was later discovered [5]. Thanks to the advances in modern channel codes, very good practical Slepian-Wolf codes have been proposed that approach the theoretical performance bound. As for lossy DSC, Wyner and Ziv first studied one of its sub-problems, lossy source coding with side information (SI) at the decoder (also known as Wyner-Ziv coding, WZC) [6][7]. It is shown that when the encoder does not have access to the SI, there is generally a small rate loss when compared to conventional source coding, but in some specific cases, the rate loss can be zero. Practical WZC usually employs quantization plus SWC. Some good results have been reported in the literature with capacity approaching performance bound on ideal sources. The reader is referred to [8] for a comprehensive review on DSC.

Despite the promising theoretical results of DSC and the successful coding practices on ideal sources, there is still a large performance gap when distributed coding is applied to real-world signals such as videos. This is because there is a significant difference between DSC and conventional source coding: in DSC, the decoder has no or very limited access to the source statistics to be decoded. This leads to the following chicken-and-egg dilemma: 1) to achieve better coding efficiency, it is desired that the knowledge about the source statistics is as accurate as possible before decoding; and 2) perfect knowledge about the source statistics is only available after the source is decoded. Existing DSC schemes usually utilize past statistics. However, most real-world sources are highly non-stationary. Current statistics could differ a lot from the past. DSC based on such inaccurate statistics usually

This research is supported in part by NSF grant CNS-0423386. The material in this paper was presented partially in the SPIE International Conference on Visual Communications and Image Processing (VCIP), San Jose, CA, Jan. 2008 [1], and partially in the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2009 [2].

Wei Liu, Lina Dong and Wenjun Zeng are with the Department of Computer Science, University of Missouri, MO 65211, USA. (e-mail: weiliu@mizzou.edu, linadong@mizzou.edu, zengw@missouri.edu).

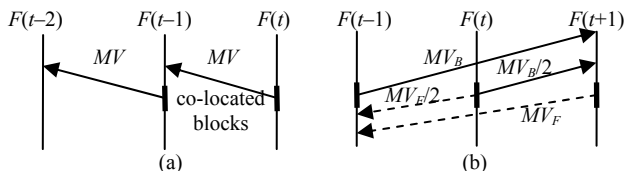


Fig. 1. 1-D illustration of (a) motion extrapolation and (b) motion interpolation at the decoder side

suffers considerable performance loss.

Let's take a closer look at DVC. Most up-to-date DVC schemes are based on the WZC architecture, so they are also called Wyner-Ziv video coding (WZVC). To perform Wyner-Ziv decoding of the encoded bits, the decoder has to generate SI first. The quality of the SI is critical because it determines the performance bound of the WZC. For WZVC, the SI is usually generated from MCP based on previously decoded frames, thus it is important to estimate the motion accurately in order to get better SI. However, motion estimation (ME) is the most computationally-intensive task in video coding. Therefore in most WZVC applications, ME is shifted from the encoder side to the decoder side. In decoder-side ME, the decoder does not have access to the current frame. This limits the accuracy of the estimated motion vectors (MV), and consequently, more bits are needed to reconstruct the current frame. This bit rate difference is referred to as the video coding loss in [9]. How to minimize the video coding loss has been the bottleneck in improving the coding efficiency in WZVC.

In [10], several approaches for decoder-side ME are reviewed and their performances are compared. A typical approach is temporal domain motion extrapolation or interpolation. For example, motion extrapolation first performs forward ME from frame $F(t-2)$ to $F(t-1)$; then for a block in $F(t)$ ², the MV of the co-located block in $F(t-1)$ is used to find its motion-compensated correspondence in $F(t-1)$. A 1-D example of motion extrapolation is illustrated in Fig. 1(a). Motion interpolation (Fig. 1(b)) is similar, but employs both forward and backward MVs for the prediction.

Here we can see that, without any knowledge about the current frame, motion extrapolation or interpolation basically assumes that objects move at a constant speed and that the estimated MVs reflect true motion, which is an over-simplification to the reality. As a result, the estimated SI is usually not a good approximation to the current frame. Analytical results in [9] suggest that WZVC could fall 6dB or more behind conventional video coding in the worst case due to inaccurate MCP.

Realizing this, researchers have proposed various methods to facilitate decoder-side motion search. For example, in [11], the encoder transmits the CRC (cyclic redundancy code) of the quantized symbols to aid the decoder in the determination of motion through Viterbi decoding. Similar approach is also found in [12], where robust hash codes are used instead of CRC. The drawback of such approaches is that overhead bits

are required to be transmitted. Backward channel-aware WZVC is proposed in [13], where the decoder feeds back several MVs as candidates for the encoder to choose from. However, real-time communication between the encoder and the decoder might not be feasible in some WZVC applications.

Approaches involving iterative motion refinement based on partially-decoded results can be found in [1][14]–[17]. Such approaches can be further categorized into two categories. In the first category [14][15], an initial estimation of the SI is first generated, then the first-pass Wyner-Ziv decoding is performed on the whole frame, leaving whatever decoding error as is in the partially-decoded frame. Then based on this result, inter-frame ME is carried out for the second-pass Wyner-Ziv decoding. The difficulty of such approaches lies in the control of bit-error-rate (BER) of the first-pass decoding. If the BER is too high, the ME result based on the partially-decoded frame might be even worse than the initial SI; on the other hand, if the BER is low, there might not be a significant rate difference when compared to the error-free decoding, since the BER curves for the channel codes employed in WZC are usually very steep.

The other category of approaches are presented in [16][17], as well as our previous work [1], where the current frame is decoded progressively in the resolution or the quality dimension. Once a low-resolution or low-quality version of the current frame is reconstructed, inter-frame ME is performed with respect to previously decoded frames to refine the motion field, and the refined motion information is employed in the Wyner-Ziv decoding of the next resolution or quality level of the current frame.

It can be seen that all of the above works rely on partial access to current frames in certain ways. This is an intuitive solution because motion is a highly non-stationary signal, both spatially and temporally, and it is always helpful to have a partial observation of the current frame in ME. However, most existing approaches are somewhat ad hoc. An in-depth understanding of their theoretical performance is warranted. In this paper, we will focus on the multi-resolution motion refinement (MRMR) approach, where low-resolution version of the current frame is iteratively decoded, based on which motion is refined. Theoretical results are provided based on power-spectrum analysis. We show that MRMR significantly improves the quality of SI and consequently, the coding efficiency of WZVC. Furthermore, we observe that in conventional ME, the resulting MVs are sent as overhead bits. Hence the motion accuracy and the bit-rate consumption of the MVs need to be jointly considered. However, in decoder-side ME, there is no such constraint. We then analyze the performance of MRMR with fractional motion search, smaller sized block matching and multiple hypothesis prediction. Results show that through extensive motion exploration, MRMR can potentially outperform conventional ME.

The rest of the paper is organized as follows. Section II provides a review on related works. In Section III, the efficiency of MRMR is modeled and compared with conventional ME and motion extrapolation. The performance

² Since most of our discussions are about the decoder, we simply use $F(t)$ to denote the decoder's observation of the t -th frame, as opposed to the original, un-coded frame.

of MRMR with extensive motion exploration is analyzed in Section IV. In Section V, a wavelet-domain codec of WZVC with MRMR is presented. Section VI concludes the paper.

II. RELATED WORKS

A. Efficiency of General MCP Coding

Girod [18] proposes a framework to analyze the R-D performance of MCP-based video coding. He shows that for a 2-D colored signal s and its MCP residual e , the following relationship holds

$$\Phi_{ee}(\omega) \approx \Phi_{ss}(\omega) \left[1 - |P(\omega)|^2 \right] \quad (1)$$

where Φ_{ss} and Φ_{ee} denote the power spectrum density (PSD) of s and e , respectively, $\omega = (\omega_x, \omega_y)$ is the 2-D spatial frequency, $P(\omega)$ is the Fourier transformed probability density function (p.d.f.) of the ‘‘motion vector displacement’’ $p(\Delta d)$, in which $\Delta d = (\Delta d_x, \Delta d_y)$ denotes the error between the estimated MV and the true MV. In the literature, a zero-mean isotropic Gaussian distribution of Δd is usually assumed. Accordingly, the rate saving that can be obtained by using MCP over intra-frame coding is

$$\Delta R = R_{MCP} - R_{intra} \approx \frac{1}{8\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \log_2 \left[1 - \exp(-\omega^T \omega \sigma_{\Delta d}^2) \right] d\omega \quad (2)$$

Eq. (2) suggests that the coding gain of MCP depends exclusively on $\text{Sigma}(\Delta d)^2$ (the variance of Δd), or the accuracy of ME. This conclusion applies to both conventional inter-frame video coding and WZVC.

The follow-up works of [18] are found in [19][20], where the efficiency of fractional-pel motion search and multiple hypothesis prediction is analyzed, respectively.

B. Efficiency of Motion Extrapolation

Li et al. [9] exploit an autoregressive model and measure the accuracy of temporal domain motion extrapolation as

$$\sigma_{\Delta d-ext}^2 = (1 - \rho_t^2) \sigma_d^2 \quad (3)$$

where ρ_t is the correlation between the motion fields in adjacent frames, and σ_d^2 is the variance of the ‘‘true motion’’.

Substituting (3) into (2) we get the rate saving using motion extrapolated side information estimation

$$\Delta R_{ext} \approx \frac{1}{8\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \log_2 \left[1 - \exp(-\omega^T \omega (1 - \rho_t^2) \sigma_d^2) \right] d\omega \quad (4)$$

The temporal motion correlation is estimated in [9] for several QCIF sequences at 30 frames per second (fps), of which the value ranges from 0.31 to 0.85. The results in [9] show that the rate-saving performance of motion extrapolation drops quickly as ρ_t decreases. For sequences with low ρ_t , motion extrapolation could fall 6 dB or more behind conventional inter-frame ME.

It is also worth noting that (3) assumes that the decoder knows the true motion field between the previous two reconstructed frames. This assumption does not hold in practice, because additional motion error will be introduced due to the inefficiency of practical ME algorithms such as block matching (as will be discussed in the next section). So (4) can be seen as an upper bound of the performance of motion extrapolation.

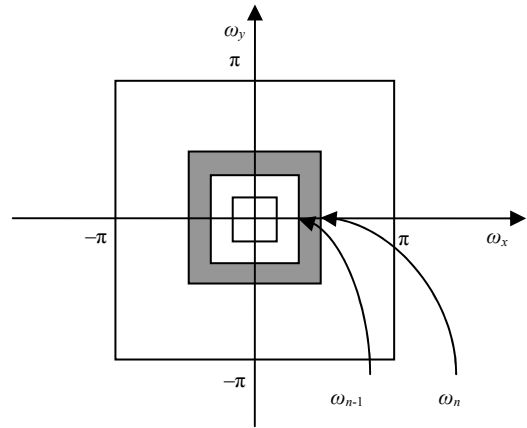


Fig. 2. Illustration of the filter bank employed in MRMR. The shaded area denotes the frequency band of the n^{th} ideal band-pass filter.

III. WYNER-ZIV VIDEO CODING USING MULTI-RESOLUTION MOTION REFINEMENT

Let’s consider the following coding paradigm of WZVC. Suppose the video frames are infinitely large and are sampled at the integer grid. The discrete time Fourier transform (DTFT) of the frame creates a periodic frequency spectrum. Due to the periodicity, we only have to transmit the spectrum in one period $|\omega_x| < \pi, |\omega_y| < \pi$. Let $0 = \omega_0 < \omega_1 < \dots < \omega_N = \pi$ be a series of frequency points, and we construct N ideal band-pass filters based on them, so that the n^{th} filter encompasses the frequency band $\{\Lambda(\omega_{n-1}) \setminus \Lambda(\omega_n)\}$ for $n = 1, \dots, N$, where

$$\Lambda(\omega_n) = \{(\omega_x, \omega_y) : \max(|\omega_x|, |\omega_y|) > \omega_n\} \quad (5)$$

and ‘‘ \setminus ’’ denotes the set difference. An illustration of the filter bank is shown in Fig. 2.

To encode the current frame, we apply the N filters to the frame and get N band-pass images without overlapping with each other in the frequency domain. The rate to transmit the whole image equals the total rate to transmit the N band-pass images.

At the decoder side, suppose at time n , the first n band-pass images have been received, based on which a low-pass version of the current frame is reconstructed. We assume this partially reconstructed image is critically sampled according to the Nyquist sampling theorem. Thus at a certain time instance, a *low-resolution* version of the current frame is available at the decoder. Now the decoder can use this low-resolution frame to get a refined estimation of the motion field and do the MCP for the next higher band-pass image.³ We are interested in how much we can benefit from this MRMR approach.

We assume the decoder employs a block matching algorithm (BMA) to refine the motion field. The BMA is carried out using the same parameters (block size, pel accuracy, etc.) at different resolution levels. To estimate the efficiency of MRMR, it is necessary to understand the motion search accuracy of a BMA, when there is only a low-resolution version of the current frame available.

³ There is an implicit assumption here that all of the band-pass images share the same motion field

A. Motion Accuracy of Decoder-side BMA

Buschmann [21] proposes an excellent model in estimating the accuracy for block-matching based motion search. In this model, a BMA can be described analytically by a series of data processing applied to the true motion field, namely estimation filtering, sampling, quantization and reconstruction filtering. The estimation filtering is introduced because a local neighborhood is usually used for the matching (which is similar to an averaging function). The larger the block size is, the smaller the bandwidth of the low-pass filter is. Sampling and quantization operations account for the fact that the estimated motion field has limited spatial resolution and amplitude precision. The sampling rate is also determined by the block size, and the quantization depends on whether or not fractional-pel accuracy motion search is used. The reconstruction filter is also a low-pass filter, representing the spatial interpolation (e.g. nearest-neighbor interpolation) of the estimated motion field.

In this case, the noise introduced to the original motion field consists of three parts: the low-pass filtering noise, the aliasing noise (due to sampling) and the quantization noise. For a decoder-side BMA, without the necessity to consider the overhead in transmitting the MVs, the motion field can be very densely sampled (e.g. using overlapped motion compensation) and finely quantized (e.g. using fractional-pel accuracy search), such that the contribution of aliasing noise and quantization noise diminishes, and the only operation that causes motion displacement is the low-pass estimation filtering.

In the Appendix, we show that under some reasonable assumption about the motion field, the variance of the low-pass filtering noise $\sigma_{\Delta d-lp}^2$ can be approximated as:

$$\sigma_{\Delta d-lp}^2 \approx \frac{2\sqrt{2} \ln \rho_s^{-1}}{\pi \omega_b} \sigma_d^2 \quad (6)$$

Here the parameter ρ_s denotes the motion correlation between neighboring pixels and is related to the spatial resolution of a video sequences; ω_b is the cut-off frequency of the estimation filter. If B is the 1-D block size used for matching, we have $\omega_b = \pi/B$. Therefore, the overall motion vector displacement of a decoder-side BMA can be modeled as a linear function of B :

$$\sigma_{\Delta d}^2 = \frac{2\sqrt{2} B \ln \rho_s^{-1}}{\pi^2} \sigma_d^2 = kB \sigma_d^2 \quad (7)$$

According to results in [21], we use $k \approx 0.005 \text{ pel}^{-1}$ for CIF sequences and $k \approx 0.01 \text{ pel}^{-1}$ for QCIF sequences.

B. Efficiency of MRMR

Now we are in the position to consider the motion search accuracy in MRMR. At time n , since the currently-available frequency band is $|\omega_x| < \omega_n$, $|\omega_y| < \omega_n$, the partially reconstructed image can be critically sampled at the rate of ω_n/π . If the decoder applies a BMA with block size B to the low-resolution image, the ‘‘effective block size’’ is π/ω_n times of that in the full-resolution image space (or the bandwidth of the estimation filter is ω_n/π times of the original). Hence we can replace B with $\pi B/\omega_n$ in (7) and get the motion search accuracy in the $(n+1)$ th level in MRMR as

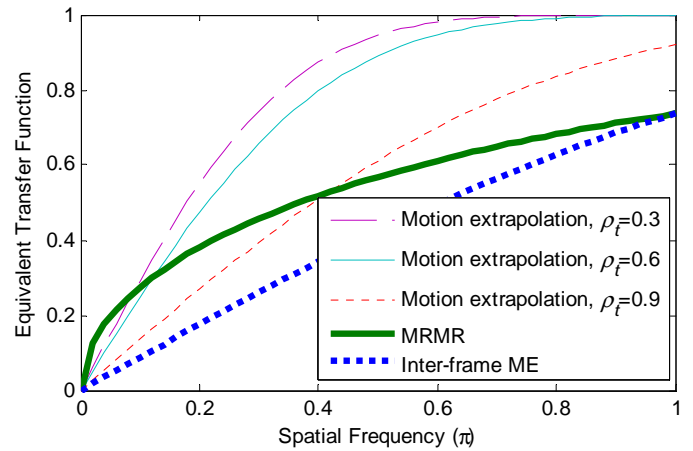


Fig. 3. 1-D equivalent transfer functions for inter-frame ME, MRMR and motion extrapolation.

$$\sigma_{\Delta d-MRMR(n+1)}^2 = \frac{\pi kB \sigma_d^2}{\omega_n} \quad (8)$$

In an ideal case, if the number of resolution levels N approaches infinite and for arbitrary n , ω_n and ω_{n+1} are close enough, for any frequency point $\omega = (\omega_x, \omega_y)^T$ in the $(n+1)$ th subband, we have $\omega_n \approx \max(|\omega_x|, |\omega_y|)$. Then the motion accuracy in (8) can be replaced with:

$$\sigma_{\Delta d-MRMR}^2(\omega) = \frac{\pi kB \sigma_d^2}{\max(|\omega_x|, |\omega_y|)} \quad (9)$$

Note that unlike the cases in inter-frame ME or motion extrapolation, the motion accuracy of MRMR is a function of ω . This can be interpreted as when higher resolution image is used for motion search, the estimated motion is also more accurate. Thus the total rate saving by using MRMR is

$$\Delta R_{MRMR} = \frac{1}{8\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \log_2 \left[1 - \exp \left(-\omega^T \omega \times \frac{\pi kB \sigma_d^2}{\max(|\omega_x|, |\omega_y|)} \right) \right] d\omega \quad (10)$$

C. Performance Analysis

In this section, performance will be compared among inter-frame ME, motion extrapolation and MRMR. For inter-frame ME, we will use (7) as a lower-bound of the MV displacement if not specifically mentioned otherwise. Therefore

$$\Delta R_{inter} = \frac{1}{8\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \log_2 \left[1 - \exp \left(-\omega^T \omega \times kB \sigma_d^2 \right) \right] d\omega \quad (11)$$

However, it should be kept in mind that for real-world inter-frame ME, aliasing error and quantization error of the MVs are generally non-negligible.

From the variance of MV displacement in (3), (7) and (9), we can see that all of them are essentially content dependent. The difference is whether the rate saving will depend on spatial motion correlation ρ_s or the temporal motion correlation ρ_t . In general, one can expect better prediction performance from motion extrapolation if the temporal motion correlation is high. However, many factors may result in the reduction of ρ_t : scene change, frame dropping, lens vibration, object moving in / out, and object moving with acceleration. As for the spatial motion correlation, except for the pixel pairs that belong to different moving objects, ρ_s is often more stable

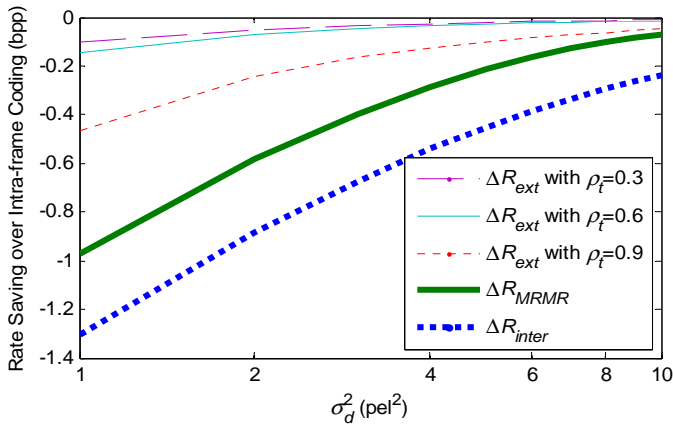


Fig. 4. Comparison of the rate saving performance among inter-frame ME, MRMR and motion extrapolation.

and higher, which is good for improving coding efficiency and rate control. It is worth noting that typically when the spatial resolution of the sequence gets higher, ρ_s also gets higher.

Eq. (1) shows that the effect of MCP can be considered as applying an equivalent spatial transfer function [18]

$$H_{MCP}(\omega) = \sqrt{1 - |P(\omega)|^2} = \sqrt{1 - \exp(-\omega^T \omega \times \sigma_{\Delta d}^2)} \quad (12)$$

to the original image (for MRMR, $\sigma_{\Delta d}^2(\omega)$ replaces $\sigma_{\Delta d}^2$). A smaller amplitude of $H_{MCP}(\omega)$ means the current frame is predicted better because there is less energy remaining in the prediction residual.

We assume $B = 4$, $\sigma_d^2 = 1$, $k = 0.01$ (QCIF) and let ρ_t take values from $\{0.3, 0.6, 0.9\}$. The frequency responses of the corresponding transfer functions are plotted in 1-D in Fig. 3. We can see that inter-frame ME always provides the best prediction. MRMR falls behind motion extrapolation in low frequency bands due to the insufficient information available. However, as more and more information becomes available at the decoder, the performance of MRMR will approach that of inter-frame ME. This means that in MRMR, the coding efficiency of the high frequency components will be closer to inter-frame ME than the low frequency components. This is an attractive property in coding of 2-D signals, because high frequency coefficients considerably outnumber low frequency coefficients.

It is also observed that for each motion extrapolation curve, there is a ‘‘critical frequency’’ at which it intersects with the curve for MRMR. This critical frequency denotes the time when there is enough information available at the decoder such that refinement of the motion field is worthwhile. A ‘‘smart’’ decoder should be able to switch the prediction mode between MRMR and motion extrapolation based on ω^* . The critical frequency can be calculated from (3) and (8) as

$$\omega^* = \frac{kB}{1 - \rho_t^2} \pi. \quad (13)$$

The critical frequency ω^* is less than $\pi/4$ even when ρ_t is as high as 0.9, meaning that it is usually safe to perform motion refinement when a quarter-resolution image (of QCIF) is decoded. Note that ω^* drops quickly as ρ_t decreases. For

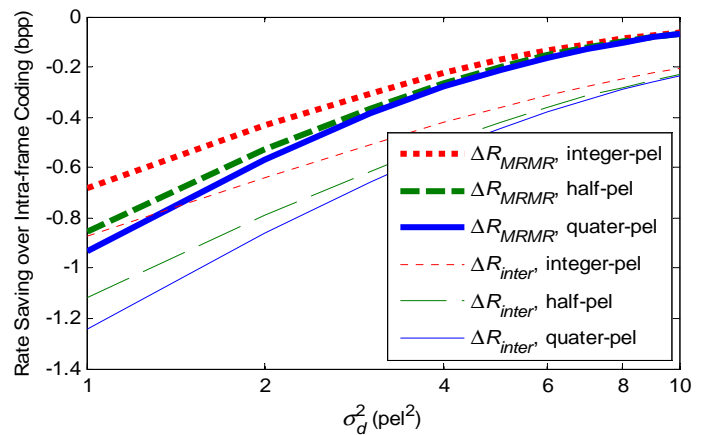


Fig. 5. Rate saving performance of MRMR and inter-frame ME using different pel-accuracy search.

sequences with medium or low ρ_t , switching between MRMR and motion extrapolation seems unnecessary, since motion extrapolation does not provide significant gain at low-frequency bands.

Next, comparison will be made on the rate saving performance (over intra-frame coding) of the three approaches. We still assume $B = 4$, $k = 0.01$ and let ρ_t take values from $\{0.3, 0.6, 0.9\}$. Numerical results are generated for ΔR_{inter} , ΔR_{MRMR} and ΔR_{ext} , and the corresponding curves are plotted in Fig. 4. A curve with a lower position means better rate saving performance.

From Fig. 4 we can see, for any approach among the three, the rate saved over intra-frame coding is more significant for more static video sequences with smaller σ_d^2 . When σ_d^2 varies from 10 to 1, MRMR can save 0.07 to 0.97 bits per pixel (bpp) over intra-frame coding. When compared with motion extrapolation, MRMR shows significant improvement. Even when ρ_t is as high as 0.9, MRMR can save 0.02 to 0.51 bpp more than motion extrapolation. It should be noted that ΔR_{ext} drops quickly as ρ_t decreases. For sequences with medium or low ρ_t , the coding gain of motion extrapolation is marginal. When σ_d^2 is as low as 1, a maximum possible saving of 0.83 bpp is observed for MRMR as compared to motion extrapolation. It is well known that for high rate coding, the rate difference at 1 bpp can be translated into 6.02 dB PSNR difference. So it can be concluded that MRMR can outperform motion extrapolation by up to 5 dB.

The comparison between inter-frame ME and MRMR shows an almost constant gap, which is averaged at 0.25 bpp when σ_d^2 varies from 1 to 10. The corresponding performance gap in terms of PSNR is 1.5 dB.

IV. MRMR WITH EXTENSIVE MOTION EXPLORATION

The analysis in the previous section is based on the assumption that both inter-frame ME and MRMR use the same settings in the BMA. However, since MRMR is performed at the decoder side, it has the advantage that the motion information does not need to be transmitted. Without this rate overhead, it is possible to exploit more advanced ME

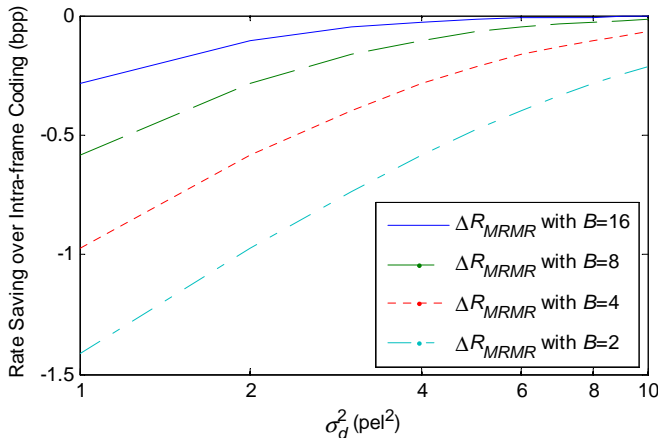


Fig. 6. Rate saving performance of MRMR with different block sizes.

techniques to extensively explore the dependency between the current frame and the reference frame(s). In the literature, better MCP can be achieved through finer fractional-pel motion search [19], using smaller block sizes or multiple hypothesis prediction [20][22]. In this section, we will analyze the performance of MRMR with these techniques.

A. MRMR with Fractional-pel Motion Search

BMA introduces quantization noise to the true motion field because the reference frame is sampled on a discrete grid. In (7) and (9), the quantization noise is not considered. In this subsection, we model the quantization error in BMAs as an additive white noise with the variance $\sigma_{\Delta d-q}^2$, hence (7) and (9) are adjusted as

$$\sigma_{\Delta d-inter}^2 = kB\sigma_d^2 + \sigma_{\Delta d-q}^2 \quad (14)$$

and

$$\sigma_{\Delta d-MRMR}^2(\omega) = \frac{\pi k B \sigma_d^2}{\max(|\omega_x|, |\omega_y|)} + \sigma_{\Delta d-q}^2 \quad (15)$$

respectively. According to [21], $\sigma_{\Delta d-q}^2$ is derived by applying uniform scalar quantization with the step size q to a random MV with the p.d.f. $p_d(d)$. We use the same settings as in [21], where $p_d(d)$ is assumed to be a generalized Gaussian distribution with the shape factor of 0.3.

Still assume $B = 4$ and $k = 0.01$, we plot the rate saving curves in Fig. 5 for integer-pel, half-pel, quarter-pel accuracy search for both inter-frame ME and MRMR. It can be seen that the rate saving using fractional-pel accuracy search in MRMR is less significant than in inter-frame ME. This can be explained from (15) that the impact of low-pass filtering noise is more significant in the MRMR case. We also conclude that in MRMR, it is more effective to perform fractional-pel motion search at the high frequency subbands where the low-pass filtering noise is less dominant.

It is also observed that when using fractional-pel search in MRMR or inter-frame ME, we can expect more gain on sequences with low motion (smaller σ_d^2). This is explained as follows. $\sigma_{\Delta d-q}^2$ is related to both q and σ_d^2 . For larger σ_d^2 , $p_d(d)$ becomes flatter and $\sigma_{\Delta d-q}^2$ approaches $q^2/12$, which is a

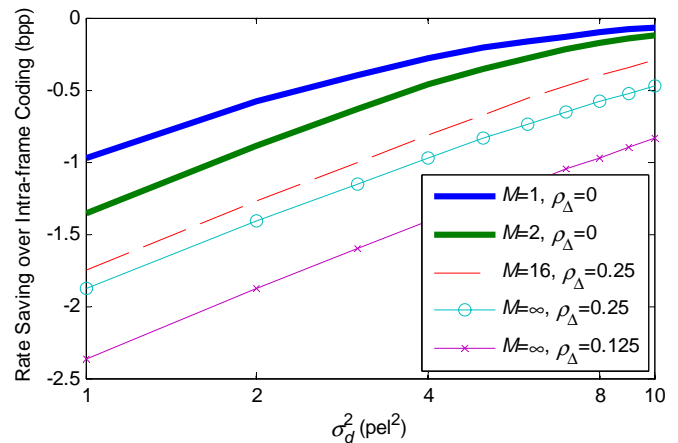


Fig. 7. Rate saving performance of MRMR with MHP.

constant and less significant than the low-pass filtering noise. While for a smaller σ_d^2 , $p_d(d)$ becomes more impulsive and $\sigma_{\Delta d-q}^2$ will approach σ_d^2 , which could be more significant than the low-pass filtering noise. Consequently, we can see the curve for ΔR_{inter} with $q = 1$ intersects with the curves of ΔR_{MRMR} with $q = 0.5$ and $q = 0.25$ when σ_d^2 is around 1.5 and 1, respectively. This means under certain complexity constraints at the encoder, MRMR might be able to outperform inter-frame ME.

The results in [9] show that improving the pel-precision does not help much in motion extrapolation. However, it is usually worthwhile to perform fractional-pel motion search in MRMR. The possible rate saving is up to 0.25 bpp if an integer-pel search is substituted by a quarter-pel search, which can be translated into 1.5 dB in PSNR gain. On the other hand, by comparing Fig. 4 (where infinitely-fine motion quantization is assumed) and Fig. 5, we can see the extra gain is marginal by employing motion search beyond quarter-pel.

B. MRMR with Smaller Block Sizes

The motivation to use smaller block sizes in MRMR is natural. By comparing (7) and (9) we know MRMR is less accurate than inter-frame ME because there is a penalty factor $\pi \left[\max(|\omega_x|, |\omega_y|) \right]^{-1}$ in its low-pass filtering noise, and the factor is always greater than 1. To compensate the penalty factor, an effective way is to use a smaller B in MRMR. In fact, For very accurate ME with small σ_d^2 , using a Taylor series expansion, (10) can be approximated as

$$\Delta R \approx \frac{1}{8\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \log_2(\omega^T \omega \sigma_{\Delta d}^2(\omega)) d\omega = C + \frac{1}{2} \log_2 B \quad (16)$$

where C is independent of B . Hence reducing B by half means an extra 0.5 bpp saving, or 3 dB gain in PSNR. This encourages the use of smaller block sizes. Certainly this conclusion also applies to encoder-side ME. However, halving the block size in each dimension also means quadrupling the number of MVs, which greatly increases the transmission overhead.

For more practical settings, we assume $k = 0.01$ and substitute different B values into (10). The corresponding

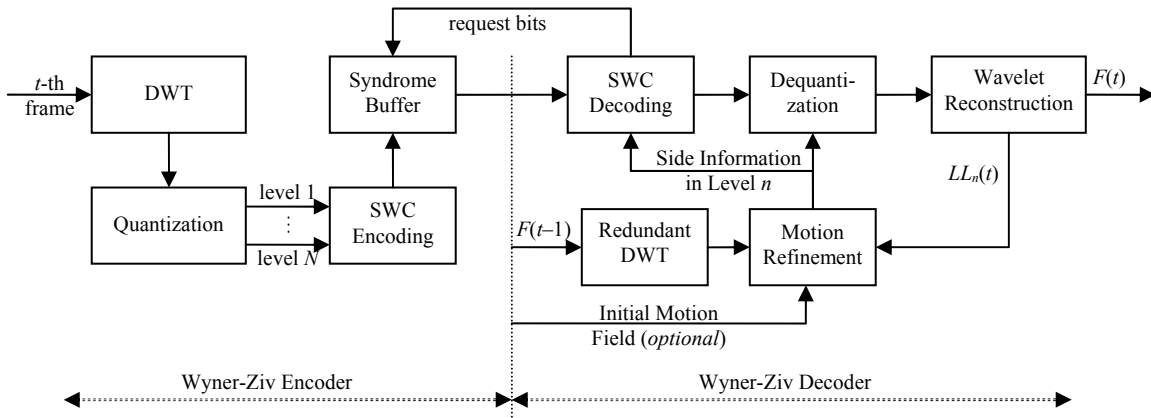


Fig. 8. Illustration of WZVC with MRMR at the decoder side

rate-saving curves of MRMR are plotted in Fig. 6. We can see that reducing the block size in BMA is effective in improving the rate-saving performance of MRMR. The $B = 2$ curve saves 0.26 bpp more than the $B = 4$ curve, which has already made up for the 1.5 dB gap with respect to inter-frame ME.

It is worth noting that block matching with a very small B is an ill-posed problem. People usually impose some smoothness constraint to make sure the derived motion field is physically meaningful. This is equivalent to applying additional inter-block low-pass filtering to the motion field, which somewhat limits the gain of reducing the block size.

C. MRMR with Multiple-hypothesis Prediction

Another effective way to improve the prediction performance in BMAs is through multiple-hypothesis prediction (MHP). One typical example is the bi-directional prediction (B pictures), where two motion vectors are assigned to the same block, with one referring to a previous frame and the other referring to a future frame. Each of the two motion-compensated blocks is called a hypothesis, and their weighted average is used for the prediction of the current block.

In this paper, we do not consider B pictures. Instead, the prediction of the current block is the weighted average of M motion-compensated blocks from previously decoded frame(s). Using the results in [20], the rate saving using MHP is

$$\Delta R_{MHP} = \frac{1}{8\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \log_2 [1 - \mathbf{P}^H \mathbf{D}^{-1} \mathbf{P}] d\omega \quad (17)$$

where we use $\Delta d_1, \dots, \Delta d_M$ to denote the MV displacements of the M hypotheses, $\mathbf{D} = [\exp(-\omega^T \Delta d_1), \dots, \exp(-\omega^T \Delta d_M)]^T$, $\mathbf{P} = E(\mathbf{D})$, and the superscript H stands for the transposed conjugate.

Following the same settings as in [22], we let the motion displacements be jointly Gaussian, each of which has the same variance (denoted as σ_Δ^2), and the correlation between any two displacements are the same (denoted as ρ_Δ). Then (17) can be simplified as (see the Appendix B):

$$\Delta R_{MHP} = \frac{1}{8\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \log_2 \left[1 - \frac{M \exp(-\omega^T \omega \sigma_\Delta^2)}{1 + (M-1) \exp(-(1-\rho_\Delta) \omega^T \omega \sigma_\Delta^2)} \right] d\omega \quad (18)$$

Note that in (18), if we set $M = 1$, the same result will be

obtained as in (2). Similar work has been done in [22], under the assumption that the M hypotheses are simply averaged, while (18) is the optimum case where the M hypotheses are Wiener filtered. Eq. (18) is for inter-frame ME, but we can simply replace σ_Δ^2 with $\sigma_\Delta^2(\omega)$ to extend it to the MRMR case. In the meantime, we would like to point out that using MHP at the encoder side will produce a lot of overhead bits – doubling the number of hypotheses also doubles the number of MVs. Therefore MHP is not widely used in state-of-the-art video coding standards. For example, in H.264/AVC P slices, only one hypothesis is allowed for each inter-predictive block.

For very accurate ME with small σ_Δ^2 , (18) can be approximated using a Taylor series expansion

$$\Delta R_{MHP} \approx \frac{1}{8\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \log_2 \left[\frac{M-1}{M} \left(\frac{1}{M-1} + \rho_\Delta \right) \omega^T \omega \sigma_\Delta^2 \right] d\omega \quad (19)$$

We can see that for a large number of hypotheses (large M), reducing ρ_Δ is equally important as reducing σ_Δ^2 ; if the displacements are mutually independent ($\rho_\Delta=0$),⁴ doubling the number of hypotheses is equivalent to reducing σ_Δ^2 by half (which means a 3 dB gain in prediction performance); however, if $\rho_\Delta > 0$, no significant gain can be obtained by increasing M when $M \gg 1/\rho_\Delta$.

Let $B=4$, $k=0.01$ and assume there is no quantization noise in the ME. We plot the rate-saving curves for MRMR in Fig. 7. Our discussions above can be well confirmed.

V. CASE STUDY

In this section, we present a wavelet domain WZVC algorithm with MRMR at the decoder side. The diagram of the proposed system is illustrated in Fig. 8.

The encoder applies an N -level wavelet decomposition to each of the frames. For each frame, $(3N+1)$ subbands are produced, namely, HL_n, LH_n, HH_n , ($n = 1, 2, \dots, N$), and LL_N .⁵ For the sake of convenience, let LL_{n-1} denote the reconstructed

⁴ ρ_Δ can certainly take negative values, but as proved in [22], $\rho_\Delta \geq (1-N)^{-1}$, so here we use 0 as the lower bound of ρ_Δ for large N s.

⁵ To accommodate the conventional notation in wavelet decomposition, here LL_N denotes the lowest frequency subband, which is different from what is used in Section 3.3.

subband from LL_n , HL_n , LH_n and HH_n , ($n = 1, 2, \dots, N$). Each of these subbands is Wyner-Ziv encoded and transmitted to the decoder at the necessary rate. To distinguish the subbands in different frames, we use notations such as $LL_n(t)$, where t is the frame index.

In our simulations, WZVC encoder decomposes each frame into 3 levels using the Daubechies 9/7 bi-orthogonal filter bank. Referring to the discussions about the ‘‘critical frequency’’ in Section III.C, 3 levels of decomposition is usually ‘‘safe’’ for most sequences. More decomposition levels only lead to marginal gain. Uniform quantization is employed to quantize the wavelet coefficients.

At the decoder side, suppose $F(t-1)$ is available,⁶ the decoding process of $F(t)$ can be described as follows:

- 1) Estimate $LL_n(t)$ from $LL_n(t-1)$ by applying MCP based on an initial motion field. The initial motion field could be initiated to be the same as $F(t-1)$, or could be simply set to zero if $F(t-1)$ is not available.
- 2) Reconstruct $LL_n(t)$ by applying Wyner-Ziv decoding.
- 3) Set $n = N$.
- 4) Refine the motion field by performing ME between $LL_n(t)$ and $LL_n(t-1)$.
- 5) Predict $HL_n(t)$, $LH_n(t)$ and $HH_n(t)$ by copying the corresponding motion-compensated coefficients in $HL_n(t-1)$, $LH_n(t-1)$ and $HH_n(t-1)$.
- 6) Apply Wyner-Ziv decoding to reconstruct $HL_n(t)$, $LH_n(t)$ and $HH_n(t)$.
- 7) Reconstruct $LL_{n-1}(t)$ based on $LL_n(t)$, $HL_n(t)$, $LH_n(t)$ and $HH_n(t)$.
- 8) Reduce n by 1. If n reaches zero, the decoding is finished; otherwise go to step 4).

To overcome the shift-variance problem in the critically-sampled wavelet domain, $F(t-1)$ needs to be transformed to the over-complete wavelet domain. Therefore, the subbands of $F(t-1)$ are all in an over-complete form, derived from non-subsampled DWT. This approach significantly improves the efficiency of MCP, at the cost of extra memory consumption in the decoder. However this strategy still fits the ‘‘simple-encoding, complex-decoding’’ principle of WZVC.

It should be noted that although iterative motion refinement is employed in our scheme, the computational complexity is even lighter than motion extrapolation, if the same BMA is used in both approaches. For example, when the motion refinement is carried out based on a half-resolution image, the number of MVs for estimation is 1/4 of that of a full-size image. For an N -level refinement, the overall complexity is $(1/4 + 1/16 + \dots + 1/4^N) < 1/3$ of the complexity of the full-resolution motion extrapolation, if the derivation of the initial motion is not considered.

Meanwhile, it is also worth noting that we actually sacrifice some performance to achieve the complexity reduction. In Section 3.3, the performance is analyzed based on the assumption that infinite levels of motion refinement are

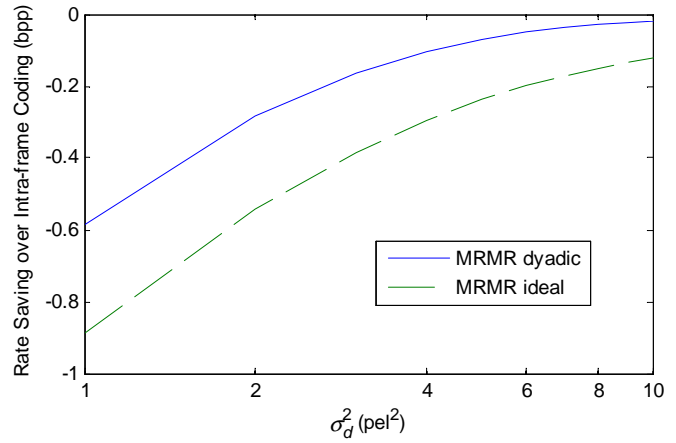


Fig. 9. Rate loss with dyadic resolution progression.

carried out. But in this codec we only allow dyadic resolution progression. Numerical results are illustrated in Fig. 9 and we can see that the additional rate loss is averaged at around 0.17 bpp, or about 1 dB loss in PSNR.

A. Benchmark Performance

In this section, we compare the performance of motion extrapolation, MRMR and inter-frame ME under the basic BMA settings. That is, sequences are encoded in IPPP fashion, each frame is divided into 8×8 blocks, one MV is searched for each block at the integer-pel precision, and only one reference frame is used. Full search is employed and the search range is $[-8, 8] \times [-8, 8]$.

In the literature, prediction performance is usually measured by the PSNR of the SI. However, we notice that the motion accuracy in MRMR is progressive (see (9)) and the SI quality is higher for high-frequency components. For the same PSNR level, the residual energy is more ‘‘compact’’ in the frequency domain. Therefore we do not directly count on the PSNR values; instead we will take a closer look at the ‘‘MSE (mean-square-error) reduction’’ performance. More specifically, we treat the residual samples in each subband as i.i.d., and calculate the rate saving of the prediction by

$$\Delta R = \frac{1}{2} \log_2 \frac{MSE_1}{MSE_0} \quad (20)$$

where MSE_1 is the MSE of the MCP, MSE_0 is the MSE of all-zero prediction (or if intra-frame coding is employed). Finally, the overall rate saving is calculated from the weighted sum of the rate saving at each resolution levels, where the weighting factors for level 1, 2 and 3 are 3/4, 3/16 and 3/64, respectively, accounting for the number of samples in each level.

Four QCIF sequences at 15 fps are used for testing: Foreman, Carphone, Mother&Daughter and News, each of which is 10 sec in length. We choose very fine quantization step sizes such that the frames are coded at very high quality (PSNR around 60dB). Thus there is no quantization error propagation and the prediction residual is almost purely due to motion mismatch. The results are shown in Table I, from which we have the following observations.

For sequences with irregular motion (small ρ_i), MRMR

⁶ Although we only describe the case of using one reference frame, it can be easily extended to multiple-frame prediction.

TABLE I. COMPARISON OF PREDICTION PERFORMANCE AMONG MOTION EXTRAPOLATION, MRMR AND INTER-FRAME ME USING BASIC BMA SETTINGS (8×8 BLOCKS, INTEGER-PEL SEARCH, ONE HYPOTHESIS FOR EACH BLOCK).

Bits per sample ^a	Foreman				CarPhone				Mother&Daughter				News			
	Lv1	Lv2	Lv3	Overall	Lv1	Lv2	Lv3	Overall	Lv1	Lv2	Lv3	Overall	Lv1	Lv2	Lv3	Overall
ΔR_{ext}	-0.04	-0.53	-1.16	-0.18	+0.06	-0.51	-1.20	-0.10	-1.10	-1.80	-2.58	-1.28	-1.38	-1.67	-2.01	-1.44
ΔR_{mrmr}	-0.51	-0.87	-1.31	-0.60	-0.55	-1.04	-1.52	-0.68	-1.23	-1.86	-2.54	-1.39	-1.64	-1.70	-1.98	-1.65
ΔR_{int}	-0.62	-1.34	-2.22	-0.82	-0.64	-1.47	-2.35	-0.86	-1.29	-2.09	-2.97	-1.50	-1.72	-2.25	-2.84	-1.85

^a Negative values with a greater absolute value means better prediction.

achieves dramatic improvement over motion extrapolation. For example, 0.42 bpp more rate saving is observed for the Foreman sequence and 0.58 bpp is observed for the CarPhone sequence. While for sequences with more consistent motion (large ρ_t) such as Mother&Daughter and News, the improvement is less significant. But MRMR still outperforms motion extrapolation by 0.11 bpp and 0.21 bpp, respectively. At high rates, the corresponding rate savings obtained by MRMR can be translated into 0.66 to 3.49 dB in PSNR gain over motion extrapolation.

When compared to inter-frame ME, MRMR suffers a small amount of rate loss, which ranges from 0.11 bpp to 0.22 bpp.

It is also observed that all of the three methods achieve more rate saving for lower frequency components. This can be explained from Fig. 3, where it is shown that the transfer functions of the three methods are essentially high-pass. They attenuate the amplitudes of low-frequency components more effectively.

However, considering the number of samples in each level, the rate saving at high-frequency bands are more important for the overall prediction performance. Thus MRMR is favorable in the sense that its motion accuracy gets refined at high-frequency bands. From Table I we can see, even for the Mother&Daughter and News sequences, MRMR is worse than motion extrapolation in level three,⁷ it still outperforms in the rest two levels. At the highest frequency bands, MRMR falls only 0.06 to 0.09 bpp behind inter-frame ME.

On the contrary, motion extrapolation is poor in predicting high-frequency components, especially for sequences with irregular motion. For the CarPhone sequence, the MCP result using motion extrapolation is even worse than the original signal in the highest subbands. As for the overall prediction performance, motion extrapolation falls as much as 0.76 bpp behind inter-frame ME, or 4.6 dB worse in PSNR.

For a visual comparison, sample residual frames from the CarPhone sequence are shown in Fig. 10. Significant improvement is observed by using MRMR.

B. Extensive Motion Exploration

The analysis in Section IV shows that, representing the motion field in greater detail will improve the performance of MRMR. On the contrary, in conventional video coding, considering the overhead in transmitting motion information, rate constrained motion estimation is usually adopted to balance the bit consumption between the motion and the prediction residual. This will potentially degrade the motion accuracy. In this subsection, we will incorporate extensive motion exploration into our MRMR framework to see whether

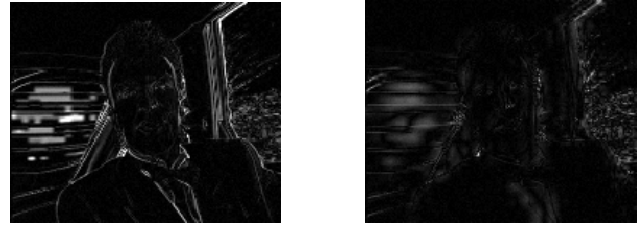


Fig. 10. Residual frame of the side information using motion extrapolation (left) and MRMR (right).

decoder-side ME can make up the gap with respect to conventional encoder-side ME, which is the major cause of performance loss in WZVC.

We enable quarter-pel search in MRMR. Pixel values at non-integer sample positions are interpolated using the same method as in H.264/AVC. Half-pel samples are derived by a six-tap finite-impulse-response (FIR) filter, and quarter-pel samples are generated from bilinear interpolation of both integer-pel and half-pel samples. The reader is referred to [23] for more details.

We use 2×2 block matching in MRMR. A smoothness constraint should be imposed to the motion field. That is, in estimating the MV for a particular block, we take into account not only the sum of absolute difference (SAD) between the current block and the candidate block, but also the sum of L_2 -norm differences between the candidate MV and its neighboring MVs. Details of this method are found in [24].

The MHP algorithm in [22] is summarized as follows. The current block first searches for the best matching block in the reference frame. This particular block is used as the initial position of all M hypotheses. The joint optimization process fixes $(M-1)$ hypotheses out of all M and searches for a new position for the remaining one. The criterion is to minimize the SAD between the current block and the average of the M hypotheses. After that, joint optimization is repeated on each of the rest of the hypotheses until the SAD converges.

According to our analysis in Section IV (C), it is desired to have the multiple MV displacements less correlated to each other. Hence we have modified the MHP algorithm as follows. We use 8 hypotheses in total, but we do not initialize all of them to the same position. Instead, only two of them are jointly optimized using the algorithm in the first stage. Then two more candidates will be added, each of which is initialized at one of the resulting places in the first stage. The second stage of optimization is carried out on the four hypotheses. Finally, the process is repeated with the addition of another four hypotheses. Such modification introduces diversity into the joint optimization process to make the MVs less correlated. Simulation also shows better results.

When combining MHP with motion smoothness constraint,

⁷ Meaning that the critical frequency here is greater than $\pi/8$.

TABLE II. COMPARISON OF THE PREDICTION PERFORMANCE BETWEEN MRMR AND H.264/AVC

Saved rate (bpp)	Akiyo	CarPhone	Container	Football	Foreman	Miss Am	M&D	News	Suzie	Salesman	Average
MRMR	-2.89	-1.70	-3.48	-0.42	-1.88	-1.46	-2.09	-2.80	-1.10	-2.44	-2.03
H.264/AVC	-2.58	-1.88	-3.04	-0.75	-2.03	-1.53	-2.12	-2.82	-1.31	-2.54	-2.06

there are issues to be addressed. Motion smoothness is defined between neighboring blocks. However, in MHP, there are multiple MVs associated with the same block, without any preferences among them. Consequently one might have to consider the motion difference of the current MV with respect to all the hypotheses of neighboring blocks, which is computationally expensive. Hence we further modify the MHP algorithm as follows. The 1st hypothesis of each block is searched for under the motion smoothness constraint (with respect to the 1st hypothesis of neighboring blocks). The resulting position is stored as the “search center” for the block. Then the multiple hypotheses are only searched in a small region around this search center. In addition, when searching around the search center, we do not consider the smoothness constraint any more to reduce the complexity.

We perform MRMR with extensive motion exploration on the 10 QCIF sequences listed in Table II and compare with H.264/AVC. For the latter, the prediction results are generated by the reference software JM13.2 [25] using the baseline profile (quarter-pel motion search, variable block size with the minimum being 4×4, at most one hypothesis for each block. The first 100 frames are tested, and for each frame (except the first frame), only the previous frame is used for MCP in both cases.

From Table II we can see, with extensive motion exploration, the prediction performance of MRMR can be very close to that of state-of-the-art inter-frame coding (only 0.03 bpp difference on average), *without any rate overhead in transmitting the motion information*. We can also see that typically, if MCP is effective for the sequence, where the achievable rate saving is higher than 2 bpp for both methods, MRMR is as good as, or even better than H.264/AVC; otherwise MRMR still suffers a performance gap (except for the Miss America sequence). This is partly because in MRMR, we simply set the prediction to be zero if no match is found in the ME, while in H.264/AVC, the intra-prediction mode will be triggered in such case, which also provides descent prediction.

It is worth noting that the final performance of WZVC is related to not only the quality of SI, but also other aspects such as the correlation model between the SI and the current frame and the efficiency of the Slepian-Wolf coder. But the quality of SI does provide a reliable estimate of the upper-bound of the R-D performance, especially at high rates. Certainly there is still a long way to go for WZVC to match the coding efficiency of conventional video coding. Our research shows that the gap is not insurmountable.

VI. CONCLUSIONS AND FUTURE WORKS

The bottleneck in improving the coding efficiency of WZVC is the performance of ME at the decoder side. The ME accuracy can be improved if the decoder has partial access to

the current frame. In this paper, we provide an analytical study on the R-D performance of WZVC in a particular structure, where low-resolution images are progressively decoded and used for the motion refinement. Theoretical analysis shows that MRMR outperforms motion extrapolation dramatically, and falls by only a small constant gap behind conventional inter-frame ME.

Realizing that decoder-side ME can benefit from more detailed motion information, which, if generated and transmitted by the encoder, incurs non-negligible bit overhead, we also provide theoretical analysis of the performance achieved by integrating MRMR with advanced ME techniques, including fractional-pel motion search, ME with smaller block sizes and multiple-hypothesis prediction. Results show that with extensive motion exploration, decoder-side MRMR can even outperform encoder-side ME with basic settings.

In addition, we present a wavelet domain practical implementation of WZVC with MRMR. Simulation results show its superior performance over the simple motion extrapolation approaches. When combined with advanced ME techniques, MRMR shows comparable prediction performance as H.264/AVC.

Future investigation will be carried out on estimating the efficiency of iterative motion refinement based on partially decoded frames with quality progression. We believe better understanding of the motion accuracy is the key to improve the performance of WZVC, and also has a significant impact on conventional video coding.

APPENDIX A. DERIVATION OF (6)

In [21], the autocorrelation function of a true motion field d is characterized as isotropic and exponentially-decreasing:

$$R_{dd}(\Delta x, \Delta y) = \sigma_d^2 \exp\left(-\omega_0 \sqrt{\Delta x^2 + \Delta y^2}\right). \quad (21)$$

where ω_0 is a small constant, Δx and Δy are the difference in the coordinates for any two pixels. The corresponding PSD of the motion field is

$$\Phi_{dd}(\omega) = \frac{2\pi\omega_0\sigma_d^2}{(\omega_0^2 + \omega_x^2 + \omega_y^2)^{3/2}}. \quad (22)$$

If we model the estimation filter as an ideal low-pass one, with the frequency response being

$$H_{est}(\omega) = \begin{cases} 0 & (\omega \in \Lambda(\omega_b)), \\ 1 & (otherwise) \end{cases} \quad (23)$$

then the PSD of the low-pass filtering noise is

$$\Phi_{\Delta d \Delta d}(\omega) = \Phi_{dd}(\omega) |1 - H_{est}(\omega)|^2 = \begin{cases} \Phi_{dd}(\omega) & (\omega \in \Lambda(\omega_b)), \\ 0 & (otherwise) \end{cases} \quad (24)$$

According to the Parseval's relation,

$$\begin{aligned}\sigma_{\Delta d-lp}^2 &= \frac{1}{4\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \Phi_{\Delta d \Delta d}(\omega) d\omega = \frac{1}{4\pi^2} \iint_{\Lambda(\omega_b)} \Phi_{dd}(\omega) d\omega \\ &\approx \frac{1}{4\pi^2} \iint_{\Lambda(\omega_b)} \frac{2\pi\omega_0\sigma_d^2}{(\omega_x^2 + \omega_y^2)^{3/2}} d\omega = \frac{2\sqrt{2}\omega_0}{\pi\omega_b} \sigma_d^2\end{aligned}\quad (25)$$

where the approximation is made because for relatively small block sizes, the cut-off frequency ω_b for the estimation filter is much greater than ω_0 , hence in the frequency band $\Lambda(\omega_b)$, we have $\max(|\omega_x|, |\omega_y|) \gg \omega_0$, and correspondingly, $\omega_x^2 + \omega_y^2 \gg \omega_0^2$.

On the other hand, according to the autocorrelation function in (21), the spatial motion correlation between two neighboring pixels is

$$\rho_s = \exp(-\omega_0). \quad (26)$$

Substituting (26) into (25) we get (6).

It might be an oversimplification to model the estimation filter as ideally low-pass. In fact, we also tested the numerical results when $H_{est}(\omega)$ is modeled as the Fourier transform of the rectangular window and the raised cosine window. The corresponding curves, together with the ideal low-pass filtering case, are plotted in Fig. 11. We can see that in all three cases, $\sigma_{\Delta d-lp}^2$ is nearly linear for a large range of block sizes, but with a slightly different slope. Since for the rectangular window and the raised cosine window, $\sigma_{\Delta d-lp}^2$ cannot be expressed in a closed form, (6) will be used as a good approximation throughout this paper.

APPENDIX B. SIMPLIFICATION OF (17)

Similar to the discussion in [22], we have

$$\mathbf{P} = P(\omega, \sigma_{\Delta}^2) \mathbf{1}. \quad (27)$$

and

$$E(\mathbf{D}\mathbf{D}^H) = P(\omega, 2(1-\rho_{\Delta})\sigma_{\Delta}^2) \times \mathbf{1}\mathbf{1}^T + \text{diag}(1 - P(\omega, 2(1-\rho_{\Delta})\sigma_{\Delta}^2)). \quad (28)$$

where $P(\omega, \sigma^2) = \exp(-\frac{1}{2}\omega^T \omega \sigma^2)$ is the Fourier transform of a Gaussian p.d.f., and $\mathbf{1} = [1, \dots, 1]^T$.

On the other hand, if matrix $\mathbf{C} = b \times \mathbf{1}\mathbf{1}^T + \text{diag}(a - b)$, its inverse can be written as

$$\mathbf{C}^{-1} = \frac{-b \times \mathbf{1}\mathbf{1}^T + \text{diag}(a + (N-1)b)}{(a-b)(a - (N-1)b)}. \quad (29)$$

Substitute a with 1, b with $P(\omega, 2(1-\rho_{\Delta})\sigma_{\Delta}^2)$, and insert (27)–(29) into (17) we get (18).

REFERENCES

- [1] W. Liu, L. Dong and W. Zeng, "Wyner-Ziv Video Coding with Multi-resolution Motion Refinement: Theoretical Analysis and Practical Significance," *Proc. SPIE Visual Communications and Image Processing (VCIP)*, San Jose, Jan. 2008.
- [2] W. Liu, L. Dong and W. Zeng, "Multi-resolution motion refinement with extensive motion exploration for Wyner-Ziv video coding", *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, April 2009.
- [3] B. Girod, A. Aaron, S. Rane and D. Rebollo-Monedero, "Distributed video coding," *Proceedings of the IEEE, Special Issue on Video Coding and Delivery*, vol. 93, no. 1, pp. 71-83, January 2005.

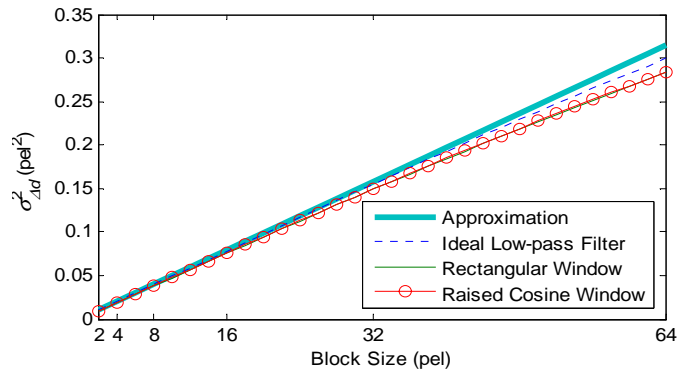


Fig. 11. Low-pass filtering noise introduced by a BMA with different block sizes.

- [4] J. D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on Information Theory*, vol. IT-19, pp. 471–480, July 1973.
- [5] A. D. Wyner, "Recent Results in the Shannon Theory," *IEEE Transactions on Information Theory*, vol. 20, no. 1, pp. 2–10, Jan. 1974.
- [6] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Transactions on Information Theory*, vol. IT-22, no. 1, pp. 1–10, Jan. 1976.
- [7] —, "The rate-distortion function for source coding with side information at the decoder—II: General sources," *Information and Control*, vol. 38, no. 1, pp. 60–80, July 1978.
- [8] Z. Xiong, A. Liveris, and S. Cheng, "Distributed source coding for sensor networks," *IEEE Signal Processing Magazine*, vol. 21, pp. 80–94, September 2004.
- [9] Z. Li, L. Liu and E. J. Delp, "Rate Distortion Analysis of Motion Side Estimation in Wyner-Ziv Video Coding," *IEEE Trans. Image Processing*, vol. 16, no. 1, pp. 98–113, Jan. 2007.
- [10] Z. Li and E. J. Delp, "Wyner-Ziv side estimator: conventional motion search methods revisited", *Proc. IEEE Int. Conf. Image Processing*, Genoa, Italy, Sep. 2005.
- [11] R. Puri and K. Ramchandran, "Prism: An uplink-friendly multimedia coding paradigm", *Proc. IEEE Int. Conf. Image Processing*, Barcelona, Spain, Sep. 2003.
- [12] A. Aaron, S. Rane and B. Girod, "Wyner-Ziv video coding with hash-based motion compensation at the receiver", *Proc. IEEE Int. Conf. Image Processing*, Singapore, Oct. 2004.
- [13] L. Liu, Z. Li, and E. Delp, "Backward Channel Aware Wyner-Ziv Video Coding", *IEEE International Conference on Image Processing*, Sept. 2006.
- [14] X. Artigas and L. Torres, "Iterative generation of motion-compensated side information for distributed video coding", *Proc. IEEE Int. Conf. Image Processing*, Genoa, Italy, Sep. 2005.
- [15] Z. Li, *New methods for motion estimation with applications to low complexity video compression*, Ph.D. dissertation, Purdue Univ., West Lafayette, IN, 2005.
- [16] J. Ascenso, C. Brites, and F. Pereira, "Motion compensated refinement for low complexity pixel based distributed video coding", *Proc. IEEE Int. Conf. Advanced Video and Signal-Based Surveillance*, Como, Italy, Sep. 2005.
- [17] B. Macchiavello, R. L. de Queiroz and D. Mukherjee, "Motion-based side-information generation for a scalable Wyner-Ziv video coder," *IEEE International Conference on Image Processing (ICIP)*, 2007.
- [18] B. Girod, "The efficiency of motion-compensating prediction for hybrid coding of video sequences," *IEEE J. Sel. Areas. Commun.*, vol. 5, no. 8, pp. 1140–1154, Aug. 1987.
- [19] —, "Motion-compensating prediction with fractional-pel accuracy," *IEEE Trans. Commun.*, vol. 41, no. 4, pp. 604–612, Apr. 1993.
- [20] —, "Efficiency analysis of multi-hypothesis motion-compensated prediction for video coding", *IEEE Trans. Image Proc.*, vol. 9, no. 2, pp. 173–183, Feb. 2000.
- [21] R. Buschmann, "Efficiency of displacement estimation techniques", *Signal Processing: Image Communication*, vol. 10, pp. 43–61, 1997.
- [22] M. Flierl, T. Wiegand, and B. Girod, "Rate-constrained multi-hypothesis prediction for motion-compensated video compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 11, pp. 957–969, Nov. 2002.

- [23] G. J. Sullivan and T. Wiegand, "Video Compression—From Concepts to the H.264/AVC Standard", *Proc. IEEE*, vol. 93, no. 1, pp. 18-31, Jan. 2005.
- [24] Y. Wang, J. Ostermann and Y.-Q. Zhang, *Video Processing and Communications*, Prentice Hall, NJ, 2001.
- [25] Available online <http://iphone.hhi.de/suehring/tml/>.